Reference Architecture Metadata Management

PostNL - Enterprise Data Management

6 december 2018 v1.0

Version control

DocVersion	Date	Version description	Distributionlist	Author	Release version
0.1	22-10-2018	Initial version	Joost van der Vlies Ernout Douqué Basten Carmio	Hugo de Gooijer	0.1
0.2	5-12-2018	Feedback Architecture Board	Ernout Douqué Basten Carmio	Hugo de Gooijer	0.9
0.3	6-12-2018	Approved Architecture Board		Hugo de Gooijer	1.0

Index metadata management for PostNL

Why

- PostNL needs metadata management to grow towards a data driven organization
- The use of metadata to describe information assets provides an array of benefits

What

- Components of metadata management
- The metadata management capabilities (purposes) will realize business value
- We will need metadata roles to perform the processes and activities
- To support the metadata processes and activities we need support IT capabilities
- Different types of metadata enable different features
- The metadata processes will realize following features to enable different types of use
- Benefits are achieved by finding the data, gaining control, improve the quality and reduced BI time

How

- The quality of metadata needs to managed as well
- The more data is shared / re-used, the more governance is required to ensure the (meta)data quality
- Categories & elements within the different metadata types that may be implemented depending on business needs
- Governance needs to be applied at the level where data is defined, designed and transformed

PostNL needs metadata management to grow towards a data driven organization

The data driven organization is:

- An organization that monitors its heartbeat through data continuously and automatically in order to control and improve their processes.

This means that the people need to work with a data driven mindset and are enabled to do so:

- Every person who can use data for improved decision making has access to this data whenever he or she needs it

These needs are enabled by metadata and requires the management of this type of data to:

- Find the data they need
- Understand that data
- Know who is responsible for that data
- Able to find the source and end users of that data
- Trust the data so they can use it without hesitation

Metadata Management:

- **Enterprise metadata management (EMM)** is the business discipline for managing the metadata about the information assets of the organization. Metadata is "information that describes various facets of an information asset to improve its usability throughout its life cycle." (*Gartner*)
- MM is a process and discipline under which metadata is collected, governed, managed and organized

Metadata

- Metadata is a type of data that digitally describes the who, what, when, where, why, and how of an organization's data, processes, applications, assets, business concepts, and/or other things of interest.

More simply we can say that metadata provides the context to the content of our digital environment

The use of metadata to describe information assets provides an array of benefits

Metadata enables discovery and retrieval

- information to data catalogues
- flexibility in searching to support cross business unit usage
- a key element in the efficient sharing of information
- the backbone of web services and interoperability
- Metadata is a precondition for self-service BI

Metadata protects investment in data

- mitigates effect of staff turnover and individual memory loss
- allows reuse and repurposing to increase return on investment
- provides documentation of data sources and quality

Metadata helps users understand data:

- Improve decision making through use of trusted data; enable process optimization with accurate data
- provides consistency in terminology and attribution
- focuses on key discerning elements of data
- helps user determine the data's fitness for use
- facilitates data transfer and interpretation by new users.

Metadata supports risk management and can limit liability

- helps prevent data from being inappropriately used or provides protection if data is inappropriately used.
- Improve quality of reporting to regulators and authorities through improved data processes and data management

Metadata is one of the solutions in the data governance toolbox

- It proves the effectiveness and added value of preventing data quality deviations by data stewardship
- It provides the insights and metrics to manage the data life cycles

Metadata reduces workload associated with questions about data

- users do not have to keep asking producers questions.

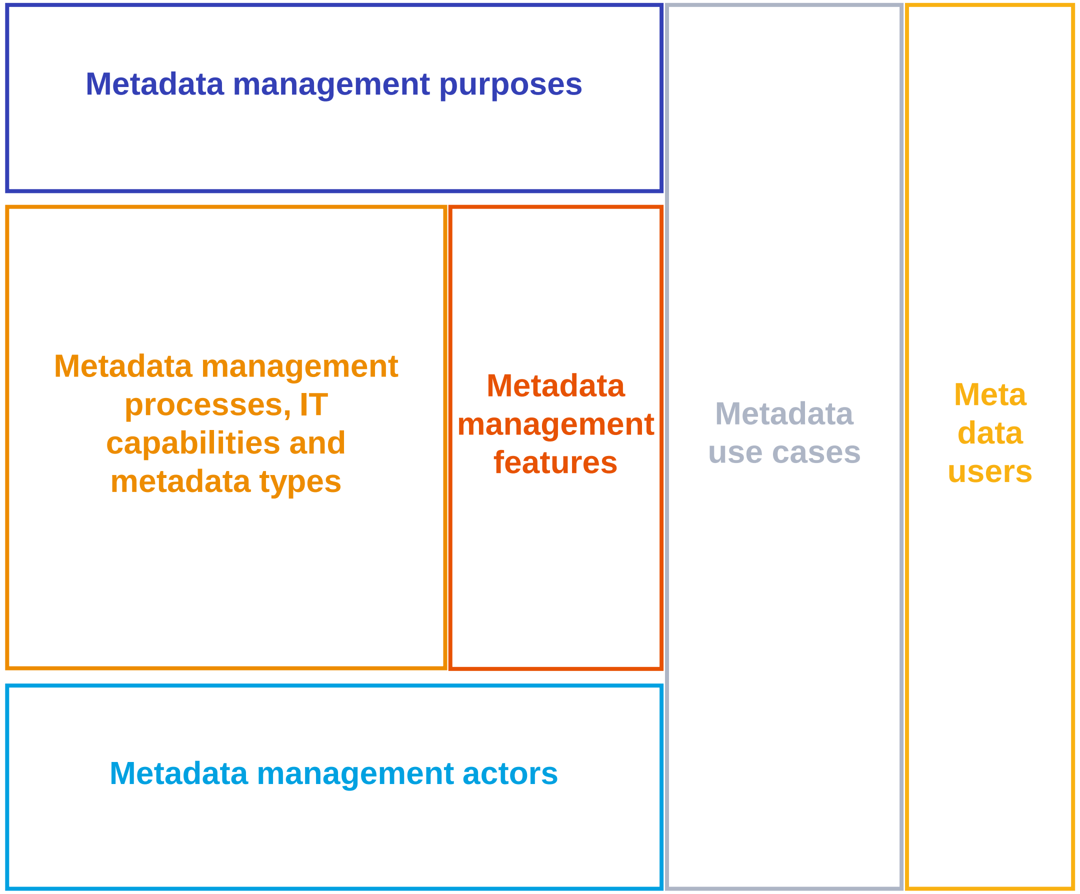
Metadata cuts overall costs

- Lower cost of data management and integration through enterprise data source mapping and enterprise access to business data definitions
- allows automation of tools which ease overall burden and cost of data population and maintenance.

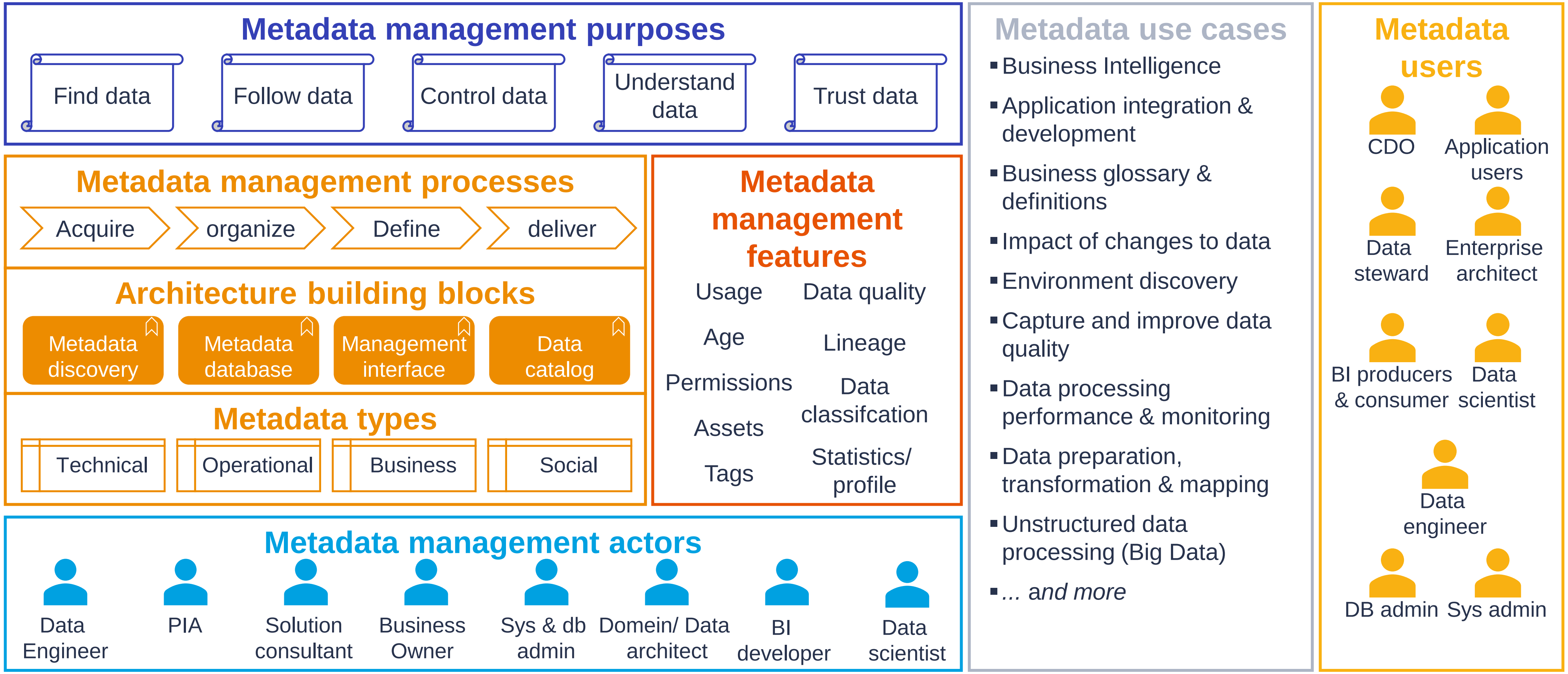
Components of metadata management

The discipline of metadata management consists of the following components:

- Knowing the users that consume the metadata
- Understanding their needs and the business purposes
- Enabling the different types of use of metadata
- Enabling the features on the metadata for each type of use
- Organizing the business processes, IT capabilities and data stores to capture, organize and use the metadata types
- Organizing governance structures and establish actors that manage the metadata


Metadata management purposes
Metadata management processes, IT capabilities and metadata types
Metadata management features
Metadata management actors
Metadata use cases
Meta data users

Metadata management framework



Metadata management purposes

- Find data
- Follow data
- Control data
- Understand data
- Trust data

Metadata management processes

Acquire → organize → Define → deliver

Architecture building blocks

- Metadata discovery
- Metadata database
- Management interface
- Data catalog

Metadata types

- Technical
- Operational
- Business
- Social

Metadata management features

- Usage
- Data quality
- Age
- Lineage
- Permissions
- Data classifcation
- Assets
- Statistics/ profile
- Tags

Metadata management actors

- Data Engineer
- PIA
- Solution consultant
- Business Owner
- Sys & db admin
- Domein/ Data architect
- BI developer
- Data scientist

Metadata use cases

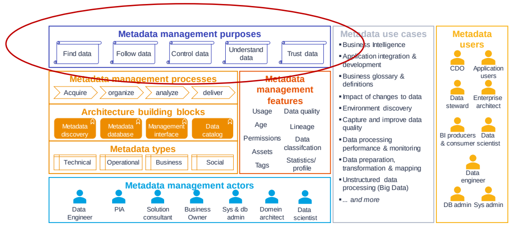
- Business Intelligence
- Application integration & development
- Business glossary & definitions
- Impact of changes to data
- Environment discovery
- Capture and improve data quality
- Data processing performance & monitoring
- Data preparation, transformation & mapping
- Unstructured data processing (Big Data)
- *... and more*

Metadata users

- CDO
- Application users
- Data steward
- Enterprise architect
- BI producers & consumer
- Data scientist
- Data engineer
- DB admin
- Sys admin

Detailed descriptions of framework blocks are explained in following slides

The metadata management capabilities (purposes) will realize business value



Metadata management purposes

Find data | Follow data | Control data | Understand data | Trust data

Metadata management processes

Acquire › organize › analyze › deliver

Architecture building blocks

Metadata discovery | Metadata database | Management interface | Data catalog

Metadata types

Technical | Operational | Business | Social

Metadata management features

Usage, Data quality, Age, Lineage, Permissions, Data classifcation, Assets, Statistics/ profile, Tags

Metadata management actors

Data Engineer, PIA, Solution consultant, Business Owner, Sys & db admin, Domein architect, Data scientist

Metadata use cases

- Business Intelligence
- Application integration & development
- Business glossary & definitions
- Impact of changes to data
- Environment discovery
- Capture and improve data quality
- Data processing performance & monitoring
- Data preparation, transformation & mapping
- Unstructured data processing (Big Data)
- *... and more*

Metadata users

CDO, Application users, Data steward, Enterprise architect, BI producers & consumer, Data scientist, Data engineer, DB admin, Sys admin

Control data

1. First we need to decide which data we want to capture the metadata for. This starts with settings priorities to control the data.

Find data

2. Then we need to find that particular data in our landscape with metadata discovery capability and capture what we find to store this in the metadata database.

Understand data

3. Next step is to understand the data for which we need a catalog and a management interface to add the business meaning

Trust data

4. We will need to analyze the data quality to be able to trust the data.

Follow data

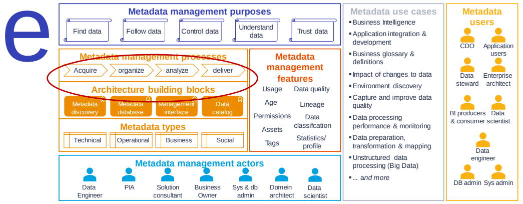
5. Lastly we will need to understand the full path from origin to recipient to get in control of the impact of data throughout the organization

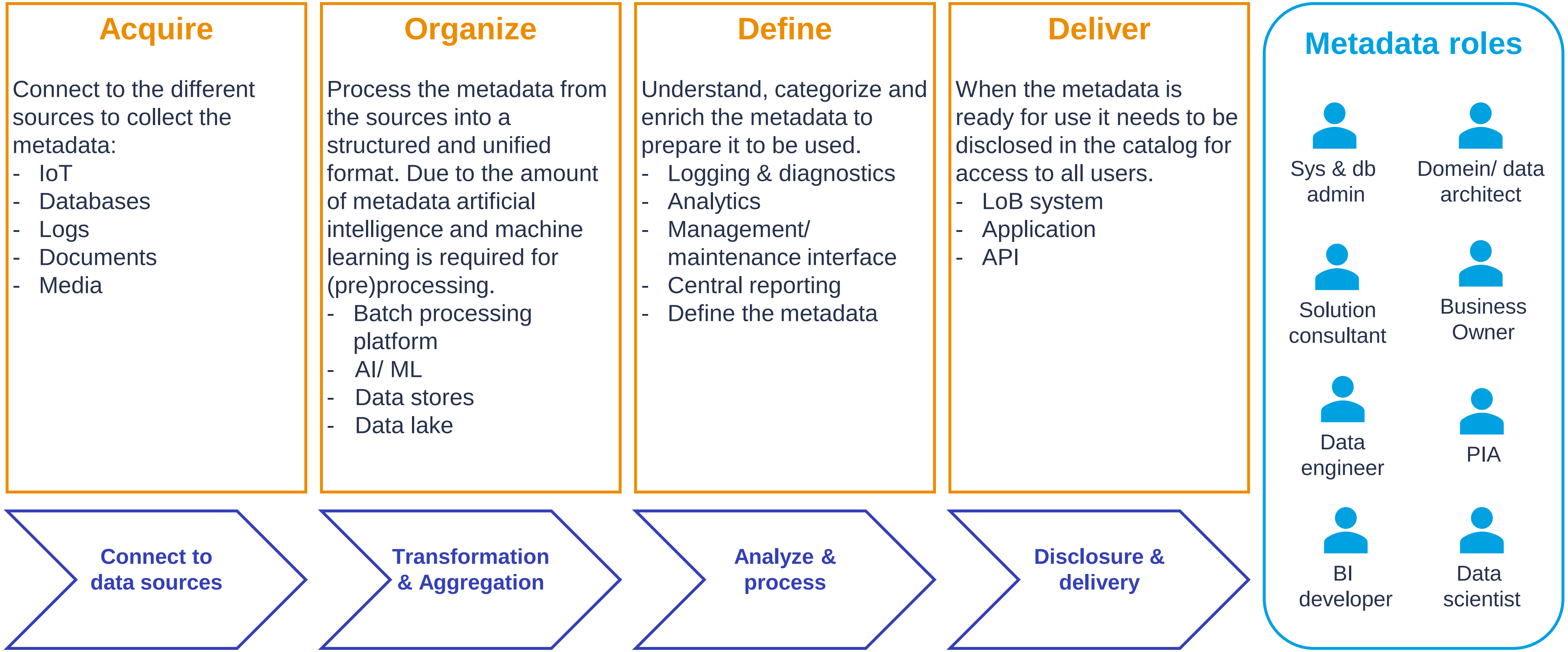
A data scientist spends a lot of time (70%) finding data, making sense of it and making it usable.

He/she needs to repeatedly discuss the data with the provider to make correct selections and interpretations.

Often, assumptions need to be made which affects the trustworthiness of outcomes and usability for business improvements.

We will need metadata roles to perform the processes and activities





Acquire

Connect to the different sources to collect the metadata:

- IoT
- Databases
- Logs
- Documents
- Media

Organize

Process the metadata from the sources into a structured and unified format. Due to the amount of metadata artificial intelligence and machine learning is required for (pre)processing.

- Batch processing platform
- AI/ ML
- Data stores
- Data lake

Define

Understand, categorize and enrich the metadata to prepare it to be used.

- Logging & diagnostics
- Analytics
- Management/ maintenance interface
- Central reporting
- Define the metadata

Deliver

When the metadata is ready for use it needs to be disclosed in the catalog for access to all users.

- LoB system
- Application
- API

Connect to data sources

Transformation & Aggregation

Analyze & process

Disclosure & delivery

Metadata roles

Sys & db admin

Domein/ data architect

Solution consultant

Business Owner

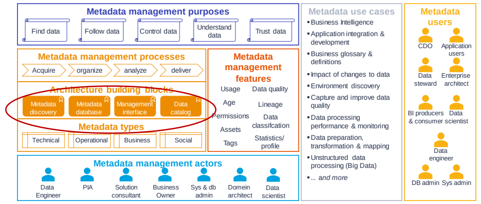
Data engineer

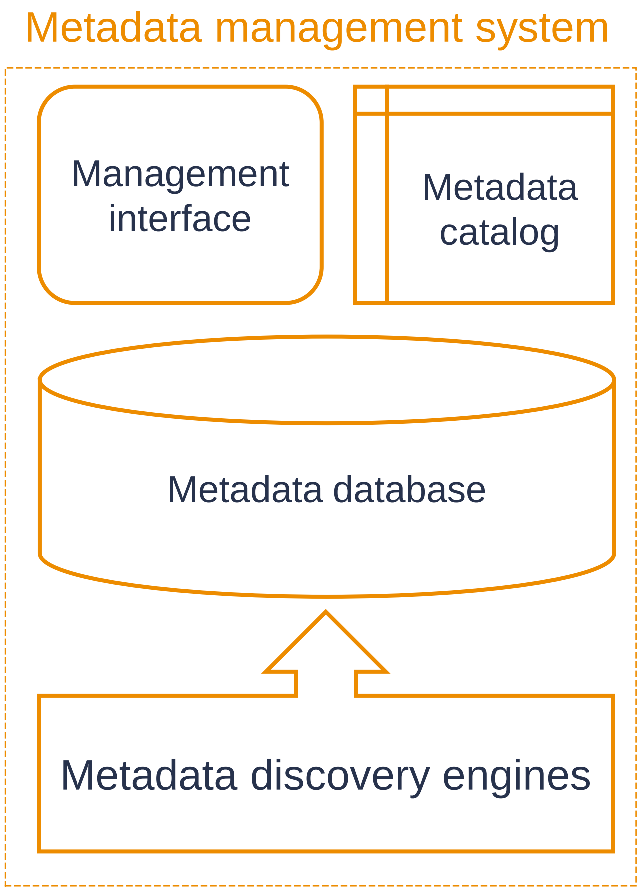
PIA

BI developer

Data scientist

To support the metadata processes and activities we need support IT capabilities





The metadata discovery engines should be able to connect with out-of-the-box connecters to most common systems containing metadata. Then discover and import that metadata into the system.
Methods required from a discovery engine:

- Reading a database management system (DBMS) catalog
- Employing ML techniques that infer metadata by analyzing file contents and structures
- Parsing ETL code

A metadata database will store all of the aggregated metadata captured from the enterprise.

- This is typically a graph database that supports various types of horizontal or vertical scaling.

A hosted data catalog and a management interface that will provide the front end to the metadata database and disclose the metadata to the PostNL. The front end should be able to

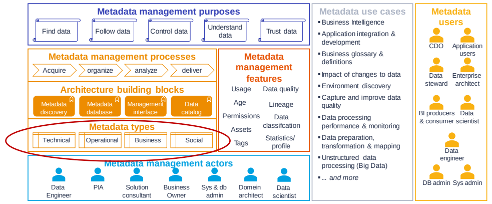
- Visualize meta data
- Support analysis through standard features
- Produce maps of data lineage
- Display the performed transformations and movements between applications

The metadata catalog contains the business glossary for PostNL with the business terms, business rules and data definitions.

It enables the business user to understand the data used in the processes and localize the data in the source and target applications.

Agile teams will be able to find and understand data faster and moreover be able to oversee the impact of changes down the chain using the management interface.

Different types of metadata enable different features



Salesforce [Account].[Name], VARCHAR(61), required yes, mandatory no

Technical *(Definitional)*

- Filetype, schemas, indexes
- Data type, length, formatting
- Data models
- Configurations
- Functions & validation rules
- References to standards
- Permissions

- Required to understand the data to be able to process it; e.g. date/time, text, numeric, integer, nullability, mandatory
- Data structure with primary & foreign keys to be able to connect data sets.
- Characteristics of the dataset and its attributes like csv comma separated, dateformat DD-MM-YYYY or MM-DD-YYYY
- Compression, encoding and encryptions

108.283 rows exported from Salesforce to HVR in 15,45s at 3-12-2018 10:05:44h with 105 exceptions

Operational *(Descriptive)*

- Output from processes & process flows
- ETL or actions on data
- Data Lineage
- Data preparation, transformation & mapping
- Performance monitoring
- Quality & audit assessments

- Find the origin of data and follow it to target applications
- Input & output of data elements with transformations
- Capture stats for each execution to monitor & enable controlled processing
  - ➢ Status, time of execution
  - ➢ Number of records/ lines read & written
  - ➢ Error counts
  - ➢ Last processed record

Business *(Descriptive)*

- Governance & stakeholders
- Reason/ purposes
- Definitions
- Business rules
- Access, privacy & security requirements
- Quality & audit requirements
- Future requirements

- Required for a business user to understand what he is looking at and the conditions how to enter data in fields
- Is data required and under what conditions
- Naming conventions for easier retrieval of data in searches
- Synonyms on business terms and abbreviations
- Availability, integrity, confidentiality and privacy classification provide directives on technical (meta)data requirements

Organization name is the legal entity name of the business partner as registered in the chamber of commerce

Social *(Descriptive)*

- Metadata about party data relationships
- User generated content
- Tribal knowledge
- Collection awareness/ communication

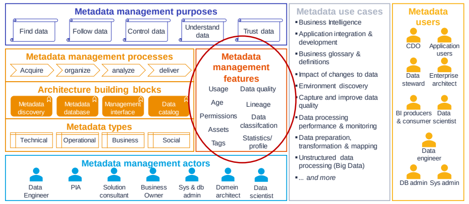
User-generated content about data can be:

- Number of views/ reads of particular data which can be used as an indicator of importance
- Social reviews and recommendations can be used to find and improve data
- Behavior on webpages
- Questions and complaints

These can provide insight into the value of the data.

Top FAQ items, based on your profile others have looked for …, areas with most outside time window delivery complaints are …

The metadata processes will realize following features to enable different types of use



Metadata management purposes
Find data
Follow data
Control data
Understand data
Trust data

Metadata management processes
Acquire
organize
analyze
deliver

Architecture building blocks
Metadata discovery
Metadata database
Management interface
Data catalog

Metadata types
Technical
Operational
Business
Social

Metadata management features
Usage
Data quality
Age
Lineage
Permissions
Data classifcation
Assets
Tags
Statistics/ profile

Metadata management actors
Data Engineer
PIA
Solution consultant
Business Owner
Sys & db admin
Domein architect
Data scientist

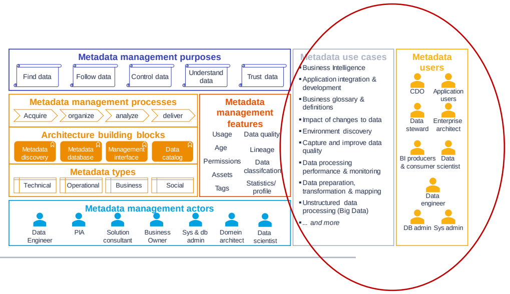
Metadata use cases
- Business Intelligence
- Application integration & development
- Business glossary & definitions
- Impact of changes to data
- Environment discovery
- Capture and improve data quality
- Data processing performance & monitoring
- Data preparation, transformation & mapping
- Unstructured data processing (Big Data)
- *... and more*

Metadata users
CDO
Application users
Data steward
Enterprise architect
BI producers & consumer
Data scientist
Data engineer
DB admin
Sys admin

Feature	Purpose	Example
Age	▪ Tracking when the data was ingested, moved or updated	▪ Updates to data ▪ <ingestion date> ▪ <copied> and <target> ▪ <last modified> ▪ <Time to live> (retention)
Assets	▪ Data on metadata assets ▪ What the asset is, and when it was created ▪ Other unique identifiers	▪ <name> ▪ <date created> ▪ <org> ▪ <location> ▪ <description>
Data quality	▪ Calculation of data validity or quality ▪ Based on factors of defined usage, and derived from sample data	▪ Generated percentage values ▪ Quality equals 0%, 25%, 50%, 100%, ▪ etc.
Data classification	▪ Marking data assets and elements with different types of classification based on internal and external standards	▪ EU GDPR-compliant ▪ <sensitive> ▪ <masked> ▪ <personal data>
Use	▪ Tracking when the data was used and for what purpose	▪ Data consumption ▪ <created_on> ▪ <accessed_on> ▪ <accessed_by>

Feature	Purpose	Example
Data Lineage	▪ Design lineage displays data flow in a design mode ▪ Real-time lineage based on history of data movement to capture real-time changes for historical reference ▪ Impact analysis of data interactions & prototyping of data movement scenarios	▪ Lineage relates to data movement ▪ Design lineage is used in architecture ▪ Real-time lineage is used in operations, in troubleshooting or for feedback to a design Model ▪ Generated model of data impacts ▪ Lineage map over time
Permissions	▪ Identifying who or what process has access to the data and who is responsible	▪ Defined permissions models ▪ <owner> ▪ <read>, <write>, <update> and <delete>
Statistics/ profile	▪ Standard statistical values, including outliers and data volume	▪ Minimum and maximum ▪ Skew and length ▪ Identified outliers ▪ Data volume
Tags	▪ Manual or generated tags identifying types of existing and custom metadata	▪ Custom tags to support categories or operations ▪ <objectID> ▪ <department> ▪ <refreshed on>

Benefits are achieved by finding the data, gaining control, improve the quality and reduced BI time



Metadata management purposes

Find data

Follow data

Control data

Understand data

Trust data

Metadata management processes

Acquire

organize

analyze

deliver

Architecture building blocks

Metadata discovery

Metadata database

Management interface

Data catalog

Metadata types

Technical

Operational

Business

Social

Metadata management features

Usage

Data quality

Age

Lineage

Permissions

Data classifcation

Assets

Statistics/ profile

Tags

Metadata management actors

Data Engineer

PIA

Solution consultant

Business Owner

Sys & db admin

Domein architect

Data scientist

Metadata use cases

- Business Intelligence
- Application integration & development
- Business glossary & definitions
- Impact of changes to data
- Environment discovery
- Capture and improve data quality
- Data processing performance & monitoring
- Data preparation, transformation & mapping
- Unstructured data processing (Big Data)
- *... and more*

Metadata users

CDO

Application users

Data steward

Enterprise architect

BI producers & consumer

Data scientist

Data engineer

DB admin

Sys admin

Roles	Use-Case Descriptions	Metadata Categories	Metadata Usage
CDO and Enterprise Architect	▪ Global metadata roles for data discovery for EIM and data governance ▪ Data discovery for specific or global domains	▪ Technical ▪ Operational ▪ Business ▪ Social	▪ Environment discovery ▪ Business glossary & definitions ▪ Impact of changes to data
Data Engineers and Data Scientists	▪ Unstructured or semistructured data discovery ▪ Big data processing ▪ AI/ML functions ▪ Feature engineering ▪ Data quality	▪ Technical ▪ Business ▪ Social	▪ Application integration & development ▪ Data preparation, transformation & mapping ▪ Unstructured data processing (Big Data)
Data Steward	▪ Implementing and maintaining data governance ▪ Data quality initiatives	▪ Technical ▪ Operational ▪ Business ▪ Social	▪ Capturing and improving data quality ▪ Environment discovery ▪ Impact of changes to data
Database Administrator and System Administrator	▪ Operational monitoring ▪ Performance metrics	▪ Technical ▪ Operational	▪ Application integration & development ▪ Data processing performance & monitoring ▪ Impact of changes to data
Operational BI and Reporting	▪ Specific business use cases ▪ LOB applications ▪ ETL, data cubes and dashboards ▪ Behavior and marketing analysis using social metadata	▪ Business ▪ Operational ▪ Social	▪ Business intelligence ▪ Application integration & development ▪ Impact of changes to data ▪ Business glossary & definitions

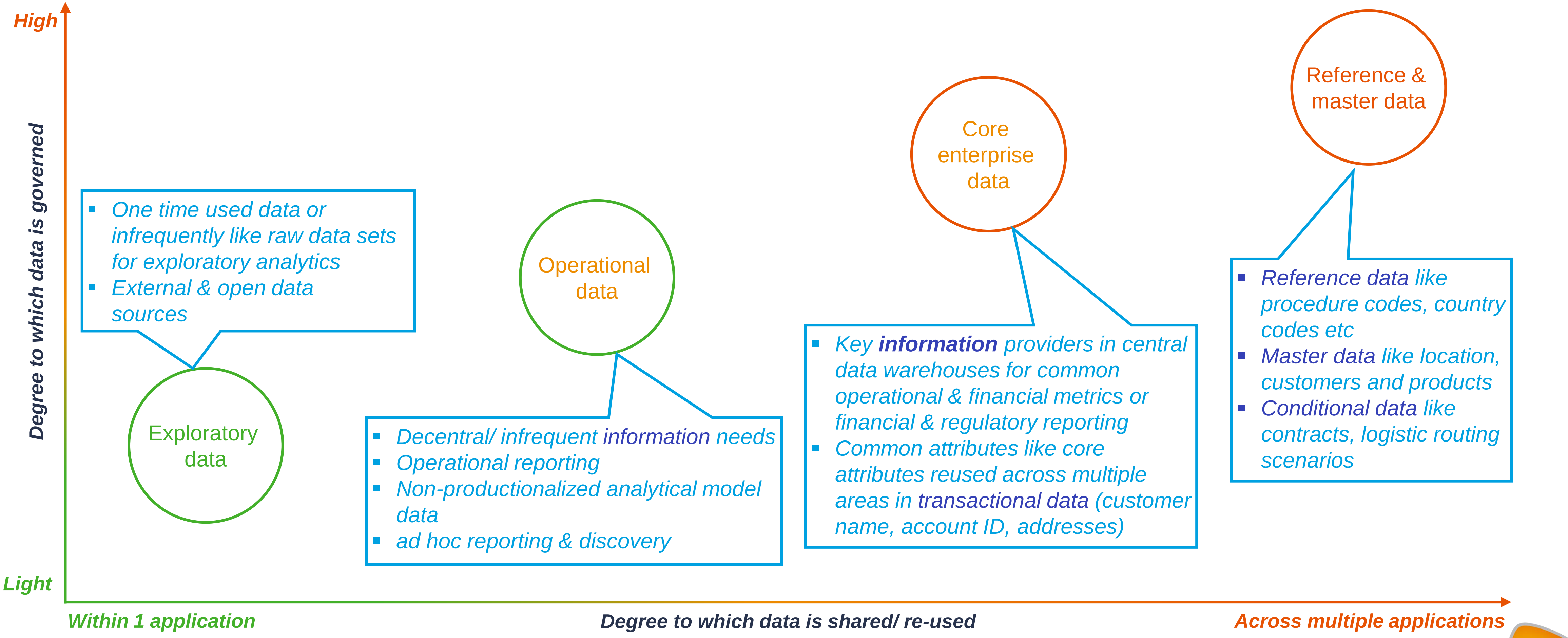
The quality of metadata needs to managed as well

Metadata is a key driver of data quality, and to support this, the metadata itself must be of high quality.

In order to ensure that quality metadata is maintained, it must be actively managed and monitored. Dashboards & Reports can be used to monitor key quality indicators:

- **Completeness**: Do definitions exist for all key data elements?
- **Accuracy**: Are current definitions correct?
  Do data types accurately represent currently implemented standards?
- **Currency/ Timeliness**: Are metadata definitions current or outdated?
- **Consistency**: Are metadata standards defined, published & implemented consistently across the organization?
- **Accountability**: Are data stewards or owners defined?
- **Integrity**: Are linkages and relationships established between critical metadata items?
- **Privacy**: Is any metadata subject to privacy restrictions?
- **Usability**: Are people actually using this metadata?

The more data is shared / re-used, the more governance is required to ensure the (meta)data quality



- *One time used data or infrequently like raw data sets for exploratory analytics*
- *External & open data sources*

- *Decentral/ infrequent information needs*
- *Operational reporting*
- *Non-productionalized analytical model data*
- *ad hoc reporting & discovery*

- *Key **information** providers in central data warehouses for common operational & financial metrics or financial & regulatory reporting*
- *Common attributes like core attributes reused across multiple areas in transactional data (customer name, account ID, addresses)*

- *Reference data like procedure codes, country codes etc*
- *Master data like location, customers and products*
- *Conditional data like contracts, logistic routing scenarios*

Categories & elements within the different metadata types that may be implemented depending on business needs

Types

Category
Elements

Technical metadata

- **Structural metadata**
  - File relationships (e.g. child, parent)
- **Technical metadata**
  - Technical table & field name
  - Data format (e.g. text, SPSS, Stata, Excel, tiff, mpeg, 3D, Java, FITS, CIF)
  - Compression or encoding algorithms
  - Encryption and decryption keys
  - Software (including release number) used to create or update the data
  - Hardware on which the data were created
  - Operating systems in which the data were created
  - Application software in which the data were created
- **Preservation metadata**
  - File format (e.g. .txt, .pdf, .doc, .rtf, .xls, .xml, .spv, .jpg, .fits)
  - Significant properties
  - Technical environment
  - Fixity information

Business metadata

- **Governance metadata**
  - Owner of the data
  - Data definition
  - Data purpose
  - Business rules
  - Data classification (BIV & Privacy)
- **Descriptive metadata**
  - Name of creator of data set
  - Name of author of the data
  - Title of document/ data
  - Object name
  - Object description
  - Object definition
  - Location of data
  - Size of data
- **Administrative metadata**
  - Information about data creation
  - Information about subsequent updates, transformation, versioning, summarization
  - Descriptions of migration and replication
  - Information about other events that have affected the files
  - Access rights metadata

Operational metadata

- **Execution metadata**
  - Whether the process run failed or had warnings
  - Which database tables or files were read from, written to, or referenced
  - How many rows were read, written to, or referenced
  - When the process started and finished
  - Which stages and links were used
  - The application that executed the process
  - Any runtime parameters that were used by the process
  - The events that occurred during the run of the process, including the number of rows written and read on the links of the process.
  - The invocation ID of the job
  - Any notes about running the process

Social metadata

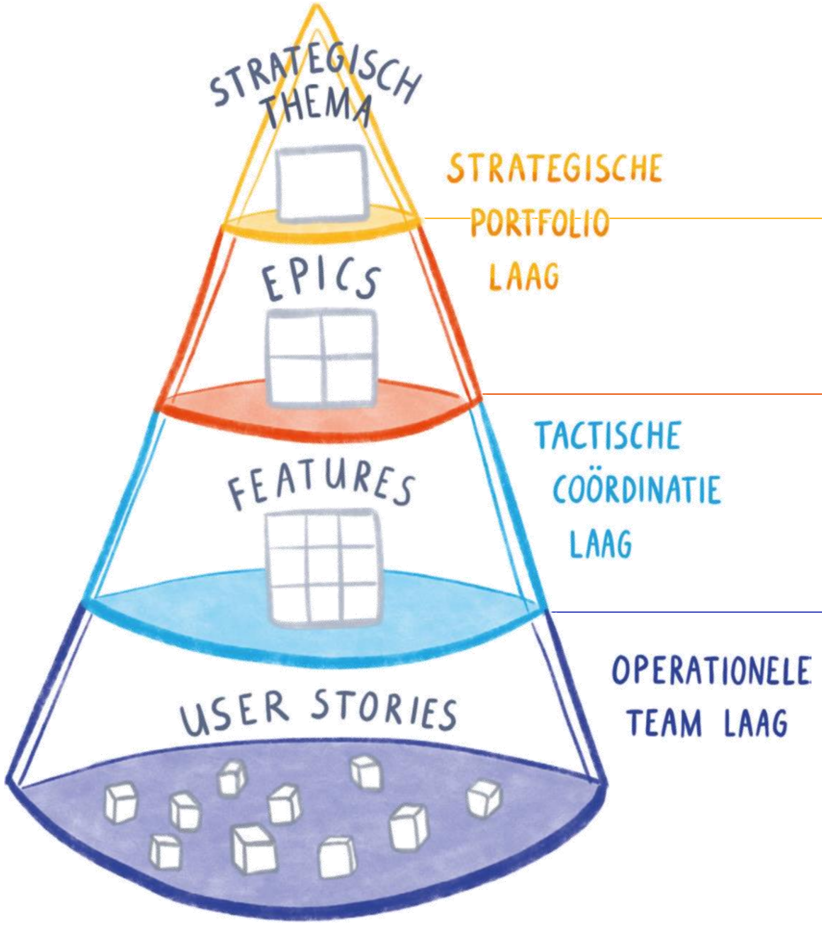
- **Use metadata**
  - Circulation records
  - Physical and digital exhibition records
  - Use and user tracking
  - Content reuse and multiversioning information
  - Search logs
  - Data tags
  - Excerpt / summary
  - URL

Governance needs to be applied at the level where data is defined, designed and transformed

- The design and definition of metadata elements that are required to use the data need to be embedded in the agile way of work

- When the data is then created and used in the business processes, the metadata will be automatically registered

- This will then enable the retrieval and re-use of the data



- **Governance metadata**

- **Descriptive metadata**

- **Structural metadata**
- **Technical metadata**
- **Preservation metadata**
- **Administrative metadata**
- **Execution metadata**
- **Use metadata**